Title
Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network

Journal
Weather and Forecasting, 22(6)

ISSN
0882-8156

Authors
Yuan, H
Gao, X
Mullen, SL
et al.

Publication Date
2007-12-01

DOI
10.1175/2007WAF2006114.1

Calibration of Probabilistic Quantitative Precipitation Forecasts with an Artificial Neural Network

HUILING YUAN* AND XIAOGANG GAO

Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California

STEVEN L. MULLEN

Department of Atmospheric Sciences, The University of Arizona, Tucson, Arizona

SOROOSH SOROOSHIAN

Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California

JUN DU AND HANN-MING HENRY JUANG

National Centers for Environmental Prediction/Environmental Modeling Center, Washington, D.C.

(Manuscript received 20 December 2006, in final form 14 May 2007)

ABSTRACT

A feed-forward neural network is configured to calibrate the bias of a high-resolution probabilistic quantitative precipitation forecast (PQPF) produced by a 12-km version of the NCEP Regional Spectral Model (RSM) ensemble forecast system. Twice-daily forecasts during the 2002–2003 cool season (1 November–31 March, inclusive) are run over four U.S. Geological Survey (USGS) hydrologic unit regions of the southwest United States. Calibration is performed via a cross-validation procedure, where four months are used for training and the excluded month is used for testing. The PQPFs before and after the calibration over a hydrological unit region are evaluated by comparing the joint probability distribution of forecasts and observations. Verification is performed on the 4-km stage IV grid, which is used as "truth." The calibration procedure improves the Brier score (BrS), conditional bias (reliability) and forecast skill, such as the Brier skill score (BrSS) and the ranked probability skill score (RPSS), relative to the sample frequency for all geographic regions and most precipitation thresholds. However, the procedure degrades the resolution of the PQPFs by systematically producing more forecasts with low nonzero forecast probabilities that drive the forecast distribution closer to the climatology of the training sample. The problem of degrading the resolution is most severe over the Colorado River basin and the Great Basin for relatively high precipitation thresholds where the sample of observed events is relatively small.

1. Introduction

Probabilistic quantitative precipitation forecasts (PQPFs) from ensemble systems provide quantitative guidance on forecast uncertainty that has the potential to improve forecast quality and utility. In contrast to a deterministic forecast, which predicts only a single outcome for precipitation quantity, an ensemble provides a discrete estimate of probability distributions across a range of precipitation rates. Timely, accurate PQPFs could provide valuable guidance for decision-makers responsible for water management, flooding warnings, and evacuations.

Yuan et al. (2005) recently used the National Centers for Environmental Prediction (NCEP) Regional Spectral Model (RSM; Juang and Kanamitsu 1994) ensemble system to produce the 24-h PQPFs. The model was run at an equivalent grid spacing of 12 km during the winter of 2002/03 over the southwest United States, and the study domain consisted of four U.S. Geological

* Current affiliation: National Research Council Associate, and NOAA/Earth System Research Laboratory, Boulder, Colorado.

Corresponding author address: Huiling Yuan, NOAA/ESRL, R/GSD7, 325 Broadway, Boulder, CO 80305-3328.
E-mail: huiling.yuan@noaa.gov

DOI: 10.1175/2007WAF2006114.1

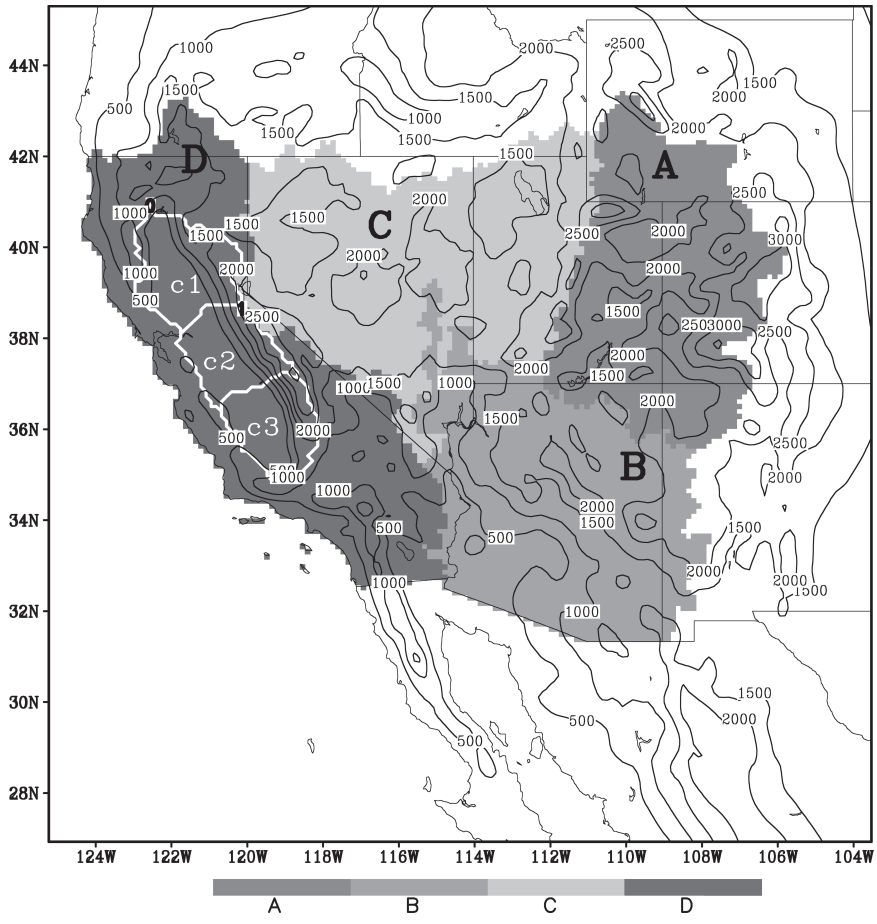

FIG. 1. The study area (163 × 172 grid points, 12-km mesh) and topography (m, contour interval 500 m). There are four USGS hydrologic regions (shaded area): the upper CO region (A), the lower CO region (B), the Great Basin region (C), and the CA region (D). Also shown are the three watersheds of the Central Valley over the CA region (boldface solid line): Sacramento (c1), San Joaquin (c2), and the Tulare basin (c3).

Survey (USGS) hydrologic unit regions: the upper and lower Colorado regions (Colorado and Arizona), the Great Basin region (Nevada), and the California region (Fig. 1). The skill of the RSM ensemble system exhibited strong dependence on geographic region and precipitation threshold. In general, the forecasts were skillful over the California region, but they showed a large wet bias over the upper and lower Colorado and the Great Basin regions.

Reduction of flow-dependent, conditional biases from ensemble precipitation forecasts is considered a necessary step to improving forecast quality and benefiting end users. Mitigation through the improvement of assimilation procedures and model formulations alone, however, poses a significant long-term challenge to the research community, especially for ensemble forecasts because of the increased dimensionality of the prediction system (e.g., Hamill et al. 2000). While systematic biases can be reduced through postprocessing (e.g., Hamill and Colucci 1997, 1998; Eckel and Walters 1998; Buizza et al. 2005), ensemble forecast systems impose additional challenges related to insufficient representation of forecast uncertainties (Gneiting and Raftery 2005) that vary by weather element, flow configuration, and ensemble formulation (for mixed-model ensembles). Calibration of ensemble systems that suffer from underdispersion or incorrect spread–error relationships (Hamill and Colucci 1997, 1998; Buizza et al. 2005; Eckel and Walters 1998) can negatively affect the ability to discriminate events (Eckel and Mass 2005). Other barriers that compromise the calibration of operational ensemble forecasts systems are insufficient training samples and ensemble sizes (Atger 2003), which are further exacerbated by the lack of independence of ensemble members (Eckel and Walters 1998). Moreover, there is little reason to hope that human

forecasters can significantly enhance ensemble systems as they are no longer able to beat current statistically postprocessed forecasts on a consistent basis (Mass 2003).

It seems clear that objective postprocessing of ensemble forecasts will remain a critical component of the forecast process. That postprocessing would include reducing errors in model-predicted fields, downscaling to scales finer than the model can resolve, or providing information on weather elements not explicitly predicted by the models (e.g., lightning, thunder, and turbulence). There are many methods of calibrating forecast systems. Among them are neural networks, which offer a proven methodology for meteorological analysis and prediction (e.g., Hsieh and Tang 1998). They have been used to improve temperature forecasts (e.g., Marzban 2003), thunderstorm forecasts (e.g., Manzato 2005), wind predictions (Kretzschmar et al. 2004), quantitative precipitation forecasts (e.g., Kuligowski and Barros 1998; Hall et al. 1999; Koizumi 1999), snowfall and snow density forecasts (Roebber et al. 2003), rainfall-runoff processes (Hsu et al. 1995), and quantitative precipitation estimation (Hsu et al. 1997; Hsu et al. 1999). However, applications of neural networks to classify (e.g., Eckert et al. 1996; Scherrer et al. 2004) ensemble members or calibrate (e.g., Mullen et al. 1998; Mullen and Buizza 2004) ensemble forecasts appear relatively limited.

In this study, an artificial neural network is applied as a postprocessor to adjust the PQPF output from the RSM ensemble. The results before and after the calibration procedure are assessed using verification measures appropriate for probabilistic forecasts of dichotomous events (e.g., Murphy and Winkler 1987). This study addresses the performance of a bias-calibration procedure, and its effectiveness and limitations on the PQPFs.

2. Model and data

NCEP includes a 45-km version of the RSM as a component of its operational short-range ensemble forecasting (SREF) system over the North American continent and the adjacent maritime zones (Du et al. 2006). The 45-km RSM provides five ensemble members: one unperturbed control run and two pairs of perturbed runs from the regional breeding method (Toth and Kalnay 1997; Du and Tracton 2001; Tracton and Du 2001). The RSM ensemble system in this study is run at an equivalent spacing grid of 12 km over the southwest United States during the cool season from 1 November 2002 to 31 March 2003 (151 days in total). The 12-km ensemble consists of 11 members: one control run and five pairs of perturbed runs that are generated by regional breeding. Finer spatial resolution is used to represent more faithfully the complex surface heterogeneity of the Southwest. The cool season is emphasized because wintertime precipitation in the semiarid Southwest supplies most of the annual freshwater, and is thereby of critical importance for the hydrology, agriculture, and water resources in the region. The model is initialized twice daily at 0000 and 1200 UTC, and dispersive lateral boundary conditions are supplied by the NCEP global ensemble forecasts at T126L28 resolution (T donates triangular wave truncation and L denotes vertical layers).

The PQPF at each model grid pixel is estimated as the fraction of the 11 members that exceed a given threshold—the so-called democratic voting method (Eckel and Walters 1998). The probability $\hat{p}_j$ for a sample j (a grid pixel during one verification time) with a precipitation rate greater than or equal to a given threshold T is calculated as the percentage of the forecast members that meet the threshold criterion; that is, $\hat{p}_j = P(\hat{x}_i \geq T)$, where $P()$ is the probability and $\hat{x}_i$ for $(i = 1, 2, \ldots, 11)$ are the model forecasts of the precipitation rate. The assumption of equally likely ensemble members to compute PQPF for the raw model output is certainly a defensible one, especially for a "classic" ensemble configuration that only considers the impact of perturbed initial conditions, and not model uncertainty (e.g., Stensrud et al. 2000).

The verification data in this study are the NCEP stage IV daily, 4-km precipitation analyses (available online at http://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage4). The 12-km RSM forecast probabilities are first interpolated bilinearly onto the stage IV 4-km grids. The interpolated PQPF data on the 4-km grids within a hydrologic unit are then compared with the concomitant stage IV estimates.

Yuan et al. (2005) document that the precipitation analyses also possess uncertainties that can significantly affect the verification scores of the RSM ensemble and, presumably, the efficacy of the calibration. In this study, however, the influence of the observational uncertainty is neglected, and the stage IV precipitation analyses are treated as a precise "ground truth." This study focuses on analyzing the effectiveness and limitations of the neural network calibration of the precipitation output from the RSM ensemble.

3. Method

A feed-forward artificial neural network is used to calibrate the RSM ensemble. The neural network (Fig. 2) includes a linear least square simplex algorithm to

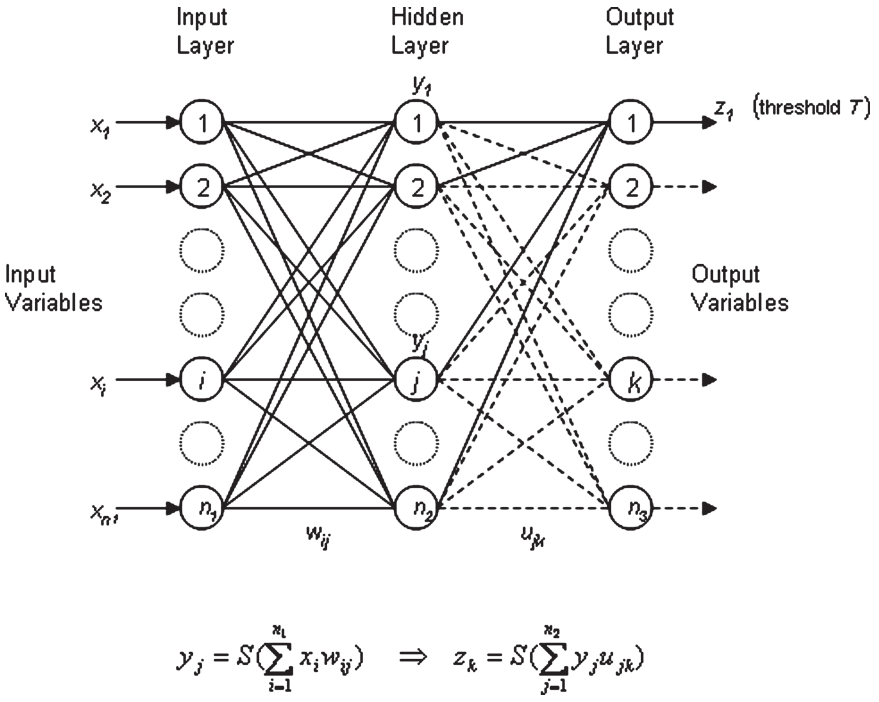

FIG. 2. Schematic of the architecture of the neural network. See text for details.

improve the search for the optimal input–output relationship in the training process (Hsu et al. 1995). The neural network consists of three layers of neural nodes, which are linked by connection weights. The neural nodes in the input layer receive a set of input variables, x_i. The nodes in the middle (hidden) layer combine the weighted values from the input nodes and modulate the summation into medium outcomes y_j through the logistic sigmoid activation function

$$S(a) = \frac{1}{1 + \exp(-a)}$$

and

$$y_j = S\left(\sum_{i=1}^{n_1} x_i w_{ij}\right).$$

Similarly, the output node calculates the final output z_k from the weighted medium outcomes from the middle layer nodes, $z_k = S(\Sigma_{j=1}^{n_2} y_j u_{jk})$. A mathematical relationship between the input and output variables (function mapping) is defined by the neural network through the optimization of the connection weights w_{ij} and u_{jk}. The strength of a neural network comes from its ability to detect complex nonlinear and unknown input–output relationships from the training samples. After training, the weights and the input–output relationship of the neural network are fixed (but they can be easily updated with additional training data through learning cycles). The input–output relationship is used to calculate the output z_k from any given input x_i.

The calibration is conducted separately at different thresholds over a hydrological unit region, for 24-h probability forecasts from the 0000 and 1200 UTC cycles, respectively. As shown in Fig. 2, the connection weights (w_{ij} and u_{jk}) are calculated through the neural network by using the input and output data. For a given threshold T, a single output variable z_k (here $k = 1$) is the target "bias free" observed probability, which is a dichotomous dataset (value is 1 for observed precipitation rates equal to or great than T; otherwise it is 0). The 18 input variables x_i (here $i = 18$) come from the RSM forecasts and include the 11 precipitation quantities predicted by the individual ensemble members and the precipitation probabilities calculated at seven thresholds centered at the given threshold T in a probability series at 15 precipitation thresholds (0.25, 1, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, and 100 mm day^{-1}). The 11 precipitation amounts are ranked, sorted, and normalized to the range [0, 1] in order to facilitate the searching relationship between the input and output data. (Note that if a "mixed physics" or a "mixed model" ensemble is used, it is crucial that the order of input data into the network not be ranked, sorted, or scrambled in order to provide the network the opportunity to assign an unequal "weighting" to the

different model configurations.) The nearest seven thresholds are chosen to compute seven probabilities in the input data. For example, at a threshold $T = 15$ mm day^{-1}, the 11 normalized precipitation rates and the seven probabilities calculated in the threshold "window" of 1, 5, 10, 15, 20, 25, and 30 mm day^{-1} are used as the input data; at a threshold $T = 5$ mm day^{-1}, the seven probabilities calculated in the threshold "window" of 0.25, 1, 5, 10, 15, 20, and 25 mm day^{-1} are used as the input data. Our convergence tests during the training revealed that this procedure led to the neural network converging in far fewer iterations [typically $O(100)$ versus $O(1000)$] without any degradation in accuracy compared to just inputting the 11 normalized precipitation values from each ensemble member. After running the neural network for the selected precipitation thresholds, the resulting calibrated probabilities were checked to ensure that the probabilities for higher thresholds were not greater than those for lower thresholds. A few outliers, on the order of 1%–3% of the grids over a hydrological region, were found that slightly missed being monotonic. They were corrected by setting the higher probability value associated with the higher threshold to the lower probability value of the adjacent lower threshold; the adjustment produced a minute drying that had an insignificant impact on the verification results.

The calibration process requires a set of training samples of the input data from the original RSM PQPF fields and the bias-free target data from the verifying observations (i.e., the stage IV daily precipitation analyses). An "objective function" is defined during the training that measures the "distance" between the calculated probability (output) and the target observed probability at a given threshold T. The neural network searches for the optimal weights to minimize the objective function and determine the input–output relationship. In this study, the root-mean-square error (RMSE) is used as the objective function at each individual precipitation threshold:

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{p}_j - o_j)^2} = \sqrt{\mathrm{BrS}}, \tag{1}$$

where $\hat{p}_j$ is the output probability calculated by the neural network for a training sample j; o_j is the target dichotomous-observed probability for the same sample defined by $o_j = 1$ if the observed precipitation rate $x_j \geq T$, otherwise $o_j = 0$; and n is the total number of training samples. Note that minimizing the RMSE is equivalent to minimizing the Brier (1950) score (BrS).

For the 5 months of 24-h precipitation forecasts over each hydrological unit region, 4 months (e.g., the 120 days from 1 November 2002 to 28 February 2003) are retained as the training period and the remaining month (e.g., the 31 days in March 2003) is used as the validation period. Cross validation is employed in which all five unique combinations of four training months and one validation month are used. Therefore, the total number of training samples n equals the number of days (~120) times the number of grid pixels in the study region, that is, the total number of available samples during the four training months. For example, about 2 million samples (16 934 pixels per day × 120 days = 2 032 080 pixels) from December 2002 to February 2003 are used to train the network over the upper Colorado region during March 2003 for a selected threshold. The n training samples are *not independent*, however, because of the spatial interdependency of 24-h precipitation and its significant in situ day-to-day correlation during the cool season. For a hydrological unit region, a set of weights for the neural network is obtained at a given threshold based on the training dataset (the RSM forecasts and threshold-dependent target-observed probabilities). Afterward, the set of achieved weights is applied to the neural network, and bias-corrected probabilities are computed for each grid pixel using the 18 input data sources from the RSM forecasts for the validation month. Results shown in the paper are a composite of the five validation months unless noted otherwise.

4. Results

a. Reliability diagrams

The neural network produces substantial improvements in the BrS and Brier skill score (BrSS; Wilks 2006) over all regions. To illustrate the performance characteristics of the bias correction, we begin by showing reliability diagrams (Wilks 2006; Jolliffe and Stephenson 2003) over the four USGS hydrologic regions for multiple thresholds (Fig. 3). The diagrams compare the conditional frequencies of the event being observed, F_i, before and after the bias correction, which correspond to the 12 discrete forecast probability levels at a given threshold T and $P_i = i/11$ ($i = 0, 1, 2, \ldots, 11$) that can be defined from the 11 RSM ensemble members. Probabilities from the neural network (NET) are not constrained to be integer values of $i/11$ ($i = 0, 1, 2, \ldots, 11$), so the NET PQPFs are binned into the closest "integer" RSM category. The conditional precipitation observation frequency is defined as $F_i = F(x_j \geq T | \hat{p}_j = P_i)$, where $F()$ is the frequency operator and x_j is the observed precipitation rate for a verifying sample j with forecast probability $\hat{p}_j = P_i$. A properly calibrated forecast has $P_i = F_i$ at any probability level i,

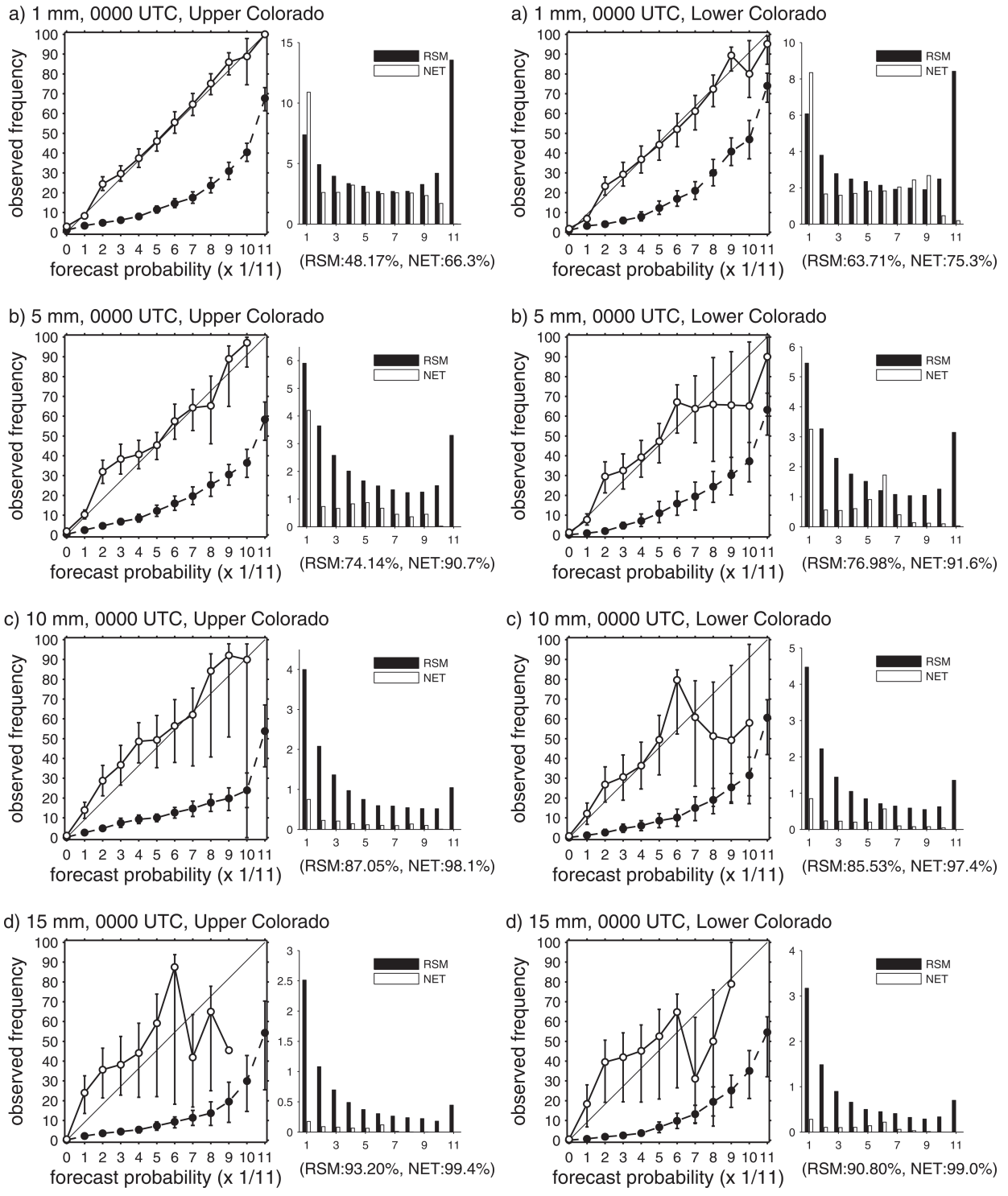

FIG. 3. (left) Reliability diagrams at four thresholds [(a) 1, (b) 5, (c) 10, and (d) 15 mm] over the upper CO region. The dashed line with black circles is for the RSM forecasts. The solid line with open circles is the NET calibration. Error bars indicate 90% confidence bounds. The histograms to the right of the reliability diagrams are frequencies (%) of each probability category (0% results shown in the parenthesis). (middle left) Same as in (left) but for the lower CO region. (middle right, facing page) Same as in (left) but for the Great Basin region. (right, facing page) Same as in (left) but for the CA region and at four thresholds [(a) 1, (b) 10, (c) 15, and (d) 25 mm].

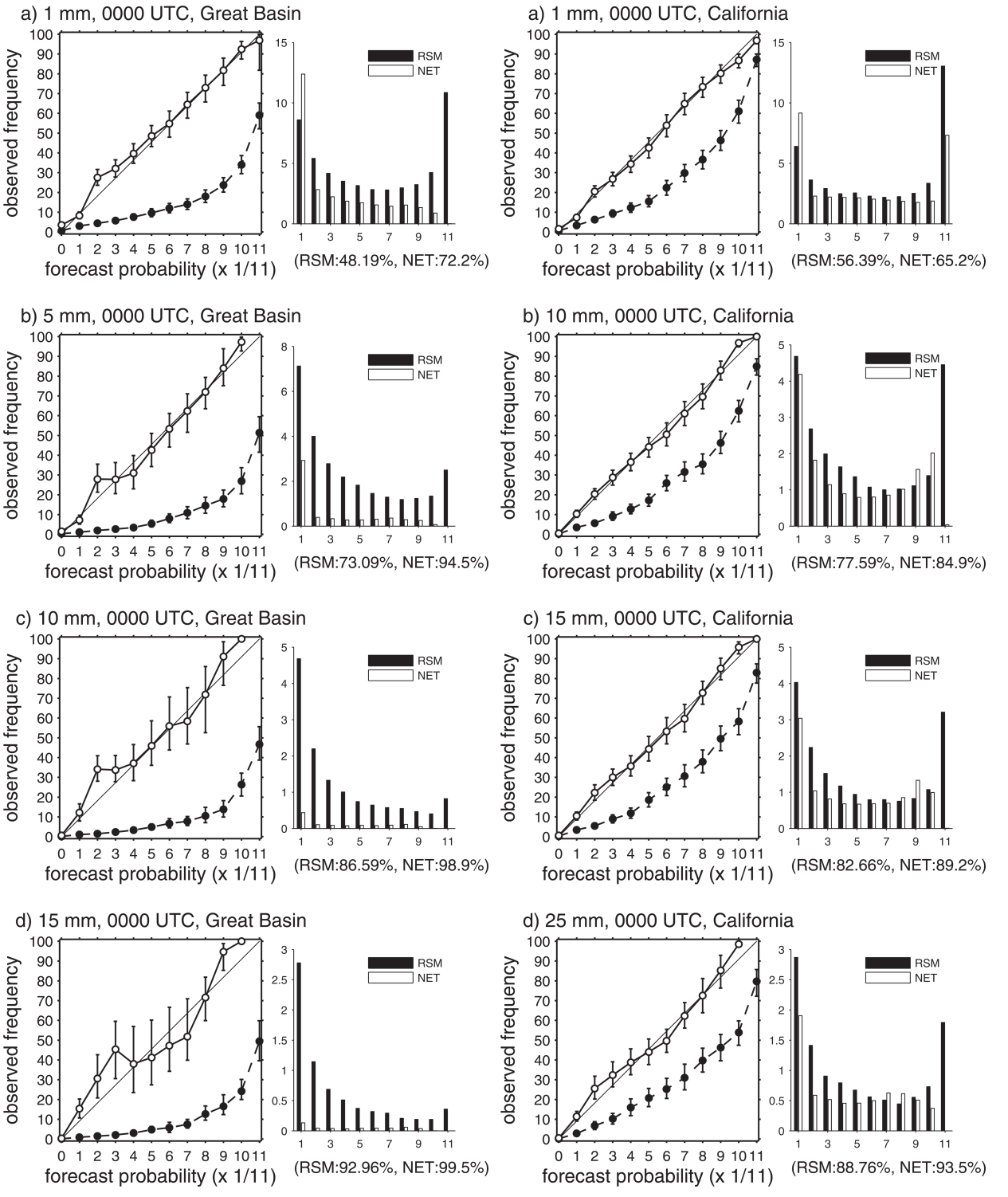

FIG. 3. (*Continued*)

so that the data points would fall on the 45° diagonal, whereas a forecast system with $P_i \neq F_i$ contains conditional biases.

The nonzero probability categories for the RSM and their 90% (95% upper, 5% lower) confidence bounds (CBs) are located below the 45° diagonal (dark circles in Fig. 3), or $P_i > F_i$ for $i \neq 0$. This indicates the presence of a pervasive wet conditional bias that affects all

thresholds over watersheds. The CBs are obtained through the bootstrap resampling method (Efron and Tibshirani 1993). The curves for the NET (open circles) are much closer to the diagonal. The degree of improvement does vary by region and threshold though. Overall, the best calibration occurs over California in terms of reliability curves being closer to the 45° diagonal and possessing tighter CBs. Lower thresholds (1 and 5 mm), in general, exhibit better calibrations than do the higher thresholds in every watershed. The NET curves for the double-digit thresholds outside of California, especially at the higher probability ranges, exhibit a sawtooth pattern and wide CBs that are characteristic of a sample size that is too small to yield a stable calibration (Wilks 2006). Consistent with the notion of a small sample, we note that the NET generates no high-confidence forecasts (probabilities of ~90% or higher) for the highest thresholds (25 mm over California, 10 or 15 mm elsewhere).

The histograms to right of the reliability diagrams give the percentage of forecasts within probability ranges that are centered about the 12 RSM categories ($i/11$, $i = 0, 1, \ldots, 11$) before (black bars labeled RSM) and after (white bars labeled NET) the calibration. The zero category (forecast probabilities between 0 and 1/22 for NET) is not shown in histograms since its magnitude is so dominant for the higher thresholds; its frequency is indicated by the percentages in parentheses. Whenever the categorical frequency is low, the uncertainty bounds (at 90% CBs) become wide, which is indicative of the estimate of the forecast probability not being robust.

Several generalities are noted for every region. A comparison of the RSM and NET histograms reveals that the neural network systematically shifts events to lower probabilities. Calibration leads to sample percentages that radically increase for nonprecipitation events, with the increase in nonprecipitation events being relatively more for higher precipitation thresholds ($\geq$10 mm day^{-1}) than for those of lower thresholds. This sample shift leads to a relative increase (correction) in the observation frequency F_i at each forecast probability level for any threshold.

Specifically, the calibration leads to an abrupt increase in the population of the 0% forecasts over the upper and lower Colorado basins and the Great Basin (Figs. 3a–c). The change comes at the expense of several higher probability levels whose populations drop to near zero after the NET correction, especially for high precipitation thresholds. This shift in the sample distribution increases the "sharpness," that is, the tendency to issue extreme forecasts as measured by the total forecast population in the outer ranks (0- and 11-member bins) or the outer two ranks (the 0–1- and 10–11-member bins). In contrast, the neural network over relatively moist California, where the RSM wet bias is not as severe, reduces the conditional bias with only minor alterations in the distribution of forecasts and a slight increase in sharpness (Fig. 3d).

b. Reliability, resolution, and uncertainty

The Brier score can be decomposed into the sum of three terms related to reliability, resolution, and uncertainty (e.g., Murphy 1973; Wilks 2006, p. 286):

$$\mathrm{BrS} = \frac{1}{n}\sum_{j=1}^{n}(\hat{p}_j - o_j)^2 = \underbrace{\left[\frac{1}{n}\sum_{i=1}^{I} N_i(P_i - \overline{o}_i)^2\right]}_{\text{Reliability}} - \underbrace{\left[\frac{1}{n}\sum_{i=1}^{I} N_i(\overline{o}_i - \overline{o})^2\right]}_{\text{Resolution}} + \underbrace{[\overline{o}(1-\overline{o})]}_{\text{Uncertainty}}, \tag{2}$$

where o_j is the observed probability with

$$o_j = \begin{cases} 1 & x_j \geq T \\ 0 & \text{otherwise} \end{cases}$$

and $\overline{o}$ is the sample climatology frequency for all verifying samples (n) with $\overline{o} = (1/n)\Sigma_{j=1}^{n} o_j$, and n is the total number of forecast–event pairs. The quantities with i subscripts denote subsample values of N_i, $\overline{o}_i$, and P_i for discrete categories in 1/11 intervals ($P_i = i/11$, $i = 0, 1, \ldots, 11$) from 0% to 100%, so $I = 12$ for this choice.

Equation (2) shows that the values of BrS, as well as the components (reliability, resolution, and uncertainty), are nonnegative ($\geq$0). The reliability term (e.g., Wilks 2006, p. 264) measures the consistency between the forecast probabilisties P_i and the conditional observation frequencies $\overline{o}_i$ at different probability subranges. Reliability equals the subsample-weighted (by N_i) squared difference between the curve and the 45° diagonal shown in Fig. 3. Reliability represents the integral bias; thus, it is termed the conditional bias. The resolution term measures the difference between the conditional observation frequencies ($\overline{o}_i$) and the climatologic frequency ($\overline{o}$) at different probability subranges, P_i. Resolution equals the subsample-weighted squared difference between the reliability curve and a horizontal line (not shown) at the climatologic frequency ($\overline{o}$) in

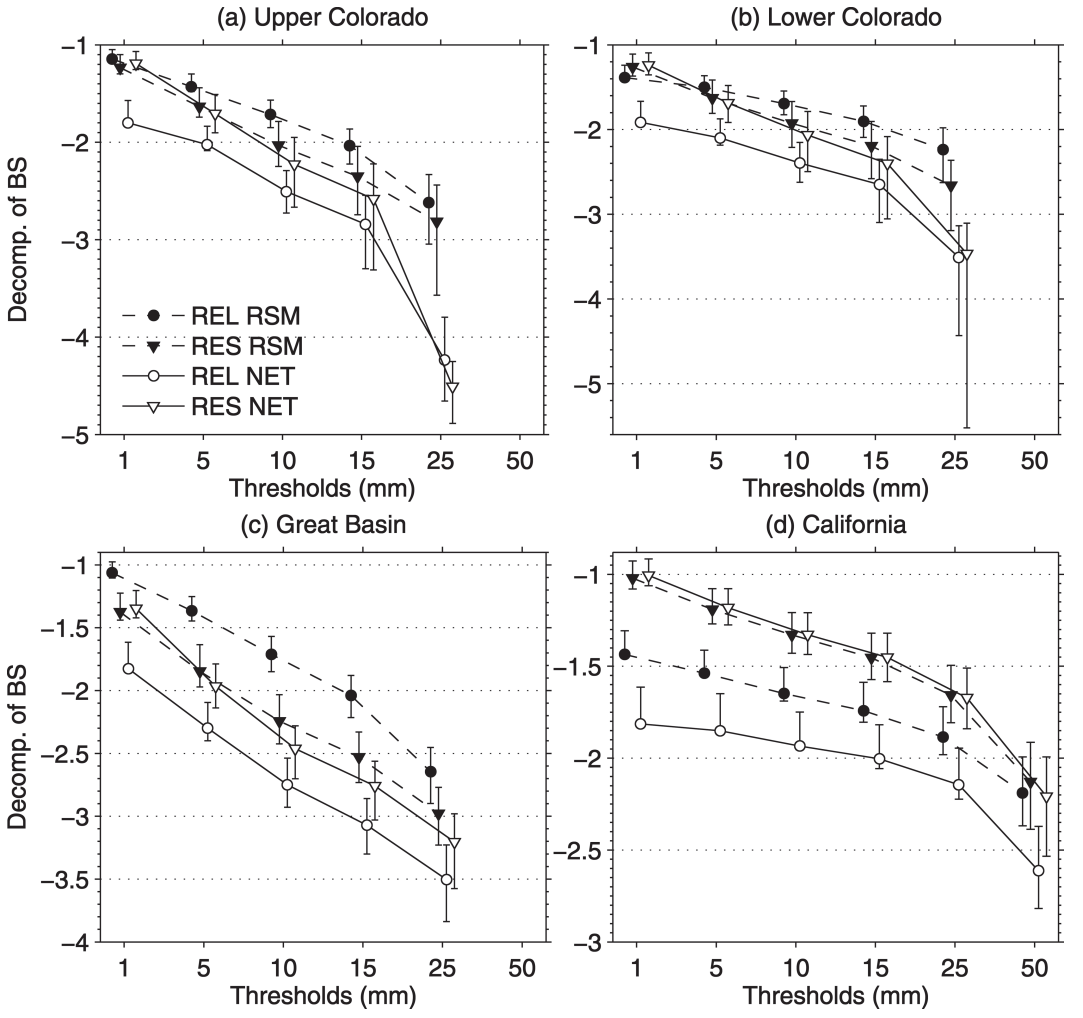

FIG. 4. Decomposition of Brier scores. Lines with circles show reliability terms. Lines with triangles show resolution terms. Dashed lines with filled symbols show the RSM. Solid lines with open symbols show the NET calibration. Shown are (a) the upper CO, (b) the lower CO, (c) the Great Basin, and (d) the CA regions. Ordinate gives the exponent for a $\log_{10}$ scale. Other three curves are slightly offset to the right of the RSM curves of the reliability terms for clarity. Error bars indicate 90% confidence bounds.

Fig. 3. A large resolution indicates that the model predicts precipitation probabilities (corresponding to a given threshold) that frequently differ from the frequency $\bar{o}$ (in contrast to the "climatologic forecast," which always predicts the same probability for a given threshold). Large resolution in the absence of a small reliability term, however, does not guarantee accurate or skillful forecasts. The uncertainty term depends solely on the sample climatology, so it does not change with the bias correction.

The BrS components at different precipitation thresholds, before and after the correction, are plotted in Fig. 4. The BrS terms have small positive values (close to zero) and vary by a few orders of magnitude over different thresholds; therefore, a $\log_{10}$ scale is used for the ordinate in Fig. 4. The 90% CBs imply that the reliability terms over the four regions at all thresholds decrease (an improvement) a significant amount after the correction, which indicates that the minimization of the BrS (as the objective function) in the training process effectively reduces the conditional bias (reliability) from the original forecasts. The neural network slightly changes the resolution term over the California region, so the NET improvement to the BrS basically results from the reduction of the reliability term in Eq. (2) with more reliable probabilities. The situation is very different over the two Colorado districts and the Great Basin, where the calibration decreases the resolution, especially at high thresholds. The degradation is not severe enough to negate the improvement in the

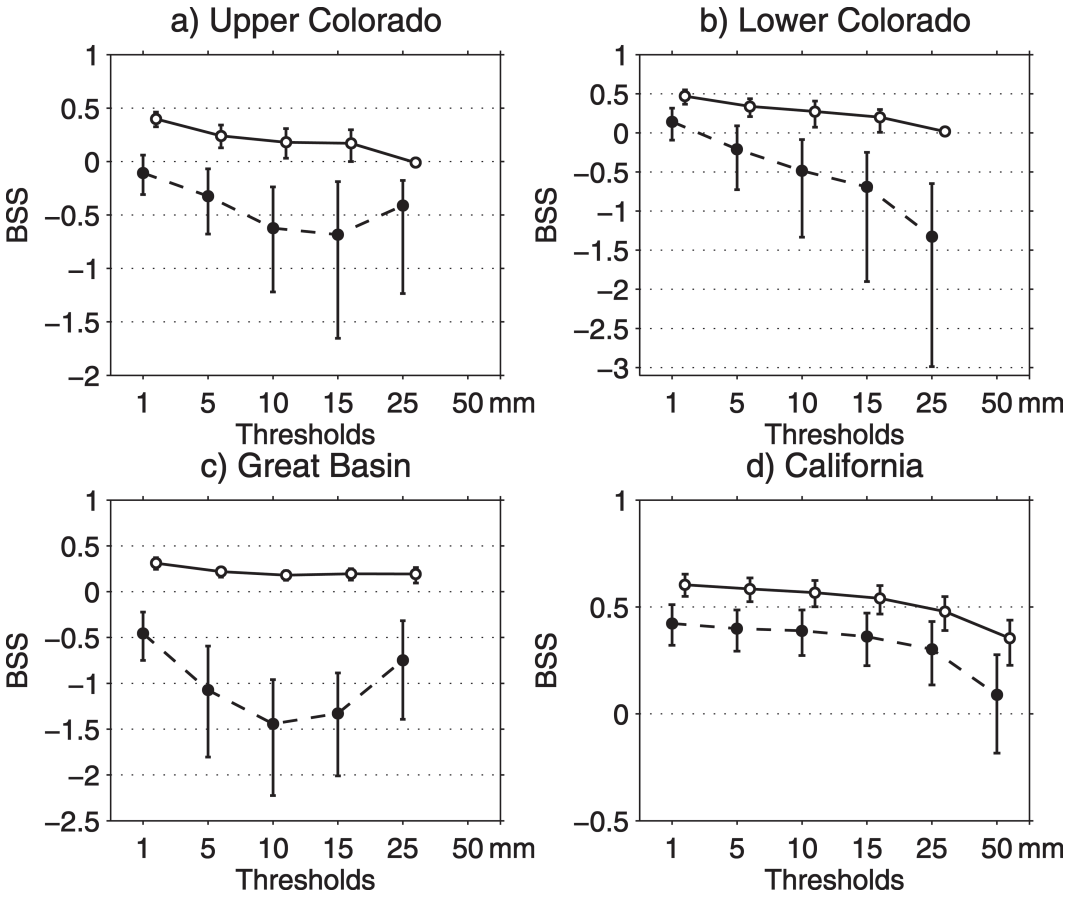

FIG. 5. Brier skill scores for the RSM forecasts (dashed line with black circles) and the NET calibrated forecasts (solid line with open circles) over four hydrologic regions: (a) the upper CO, (b) the lower CO, (c) the Great Basin, and (d) the CA. Error bars indicate 90% confidence bounds. The NET curves are slightly offset to the right of the RSM curves for clarity.

reliability, however, so the BrS values are still greatly reduced. It is clear that the bias correction over the interior watersheds comes at the cost of reduced resolution, which in terms of forecast specificity, means the NET calibration in these regions is not as effective as it is over California.

c. *BrSS*

The BrSS is commonly used to evaluate the skill of probabilistic forecasts from an ensemble model. We measure skill relative to a forecast based on the climatologic frequency of event occurrence:

$$\mathrm{BrSS} = 1 - \frac{\mathrm{BrS}}{\mathrm{BrS}_c}, \tag{3}$$

where BrS is the Brier score of the model forecasts and BrS_c is the Brier score of a forecast based on the climatologic frequency of the event occurring. In this study, the BrS_c is calculated from the 5-month precipitation observation data over each hydrologic unit. A BrSS > 0 is defined as a skillful forecast, with $\mathrm{BrSS}_{\mathrm{PERFECT}} = 1$ for a perfect forecast system (BrS = 0). The lower bound for the BrSS depends on the climatologic frequency and is

$$\mathrm{BrSS}_{\mathrm{FLOOR}} = 1 - \frac{\mathrm{BrS}_{\mathrm{WORST}}}{\mathrm{BrS}_c} = 1 - \frac{1}{\bar{o}(1-\bar{o})} \leq -3.$$

(The value of -3 occurs when $\bar{o} = 0.5$.) It is straightforward to show by substituting Eq. (2) into Eq. (3) that a positive (skillful) BrSS requires the resolution term to be larger than the reliability term.

Figure 5 shows the BrSS variation before (dashed line) and after (solid line) calibration. The neural network produces significant improvements in the BrSS for all precipitation thresholds over every USGS region. The NET curves lie well above the RSM lines, outside the respective 90% CBs for the RSM and vice versa. In fact, the lower confidence bound for the NET

TABLE 1. Monthly variations of the Brier skill score (BrSS) for the 10 mm day^{-1} threshold for the 0000 and 1200 UTC cycles (decimal numbers along the top half of the row). RSM denotes the BrSS before the calibration; NET denotes the BrSS after the calibration. Boldface numbers along the bottom half of the rows denote the rank of the BrSS for that month within each region. Cells set in italics indicate a month and forecast cycle where the ranks of the RSM and NET differ by more than one.

Month		Nov		Dec		Jan		Feb		Mar	
Region	Cycle (UTC)	RSM rank	NET rank	RSM rank	NET rank	RSM rank	NET rank	RSM rank	NET rank	RSM rank	NET rank
Upper CO	0000	−0.40	0.44	−0.69	0.21	−0.95	−0.06	−0.87	−0.06	−0.73	0.08
		1	**1**	**2**	**2**	**5**	**4.5**	**4**	**4.5**	**3**	**3**
	1200	−0.26	0.32	−1.76	−0.02	−5.95	−0.15	−2.12	−0.05	−1.58	0.11
		1	**1**	**3**	**3**	**5**	**5**	**4**	**4**	**2**	**2**
Lower CO	0000	−1.13	−0.05	−2.50	−0.10	−5.44	−0.10	−0.17	0.32	−0.31	0.37
		3	**3**	**4**	**4.5**	**5**	**4.5**	**1**	**2**	**2**	**1**
	1200	−1.22	0.04	−3.30	−0.07	−6.19	−0.06	−0.62	0.19	*−1.22*	*0.23*
		2.5	**3**	**4**	**5**	**5**	**4**	**1**	**2**	*2.5*	*1*
Great Basin	0000	−0.35	0.20	−1.42	0.20	−8.09	−0.03	−3.02	−0.05	−1.90	0.15
		1	**1.5**	**2**	**1.5**	**5**	**4**	**4**	**5**	**3**	**3**
	1200	−0.37	0.21	−2.78	0.11	−8.96	0.05	−3.98	0.13	−4.28	−0.05
		1	**1**	**2**	**3**	**5**	**4**	**3**	**2**	**4**	**5**
CA	0000	0.57	0.65	0.39	0.57	0.22	0.39	0.04	0.36	0.33	0.58
		1	**1**	**2**	**3**	**4**	**4**	**5**	**5**	**3**	**2**
	1200	0.53	0.59	0.32	0.58	−0.02	0.27	−0.17	0.36	0.13	0.54
		1	**1**	**2**	**2**	**4**	**5**	**5**	**4**	**3**	**3**

curves often lies above the upper bound for the RSM, which denotes statistical significance at the 0.25% confidence level. Over California, the network improves upon the RSM forecasts of vaguely skillful forecasts. (Note the CBs extend well below the zero line.) In the case of the three interior USGS regions, the network is able to take highly unskillful forecasts and produce skillful forecasts. The improvements seem particularly impressive for the highest thresholds (50 mm for California, 25 mm elsewhere). The RSM forecasts for heavy accumulations are either equivocal over California (CBs extend well below the zero) or unskillful to varying degrees over the interior, but the NET forecasts possess either significant skill or, in the worst case, skill no worse than that of the sample climatology.

Another notable advantage of the NET calibration is that the widths of the CBs for the BrSS are substantially reduced, especially the negative tails. Note that $\mathrm{BrSS}_{\mathrm{FLOOR}} \to -\infty$ as $\bar{o} \to 0$ or 1, since $\mathrm{BrSS}_{\mathrm{FLOOR}} \approx -1/\bar{o} \ll -3$ for $\bar{o} \to 0$ and $\mathrm{BrSS}_{\mathrm{FLOOR}} \approx -1/(1 - \bar{o}) \ll -3$ for $\bar{o} \to 1$. The condition of small $\bar{o} \to 0$ holds true for high thresholds during a 24-h accumulation interval over most semiarid regions. The absence of 90% CBs for the NET curves that extend far below the zero line indicates the individual NET forecast events that are unskillful are either sufficiently small in number or small in their level of negative skill so as to not overwhelm the net contribution from the forecasts with positive skill. In other words, the day-to-day volatility of the NET forecasts is much less than it is for the uncalibrated RSM forecasts. That is clearly not the case for the uncalibrated RSM ensemble, where some CBs dip well below −2 (e.g., lower Colorado and the Great Basin). The reduction comes at a cost, however—reduced resolution that is most acute over the interior regions.

The intraseasonal variability of calibration performance can be briefly examined by analyzing the variability of monthly skill. Table 1 gives the monthly BrSS results from the cross-validation experiments at the 10-mm threshold for the 0000 and 1200 UTC cycles. There are several points worthy of mention. First and foremost, *the calibration always increases monthly skill*, even though it may not be good enough to produce a skillful forecast for every month over the interior districts. This behavior holds true in every district and for both analysis cycles. Calibration can make moderately unskillful forecasts skillful, but it is unable to transform extremely unskillful RSM forecasts into skillful ones, such as those that exemplify the interior districts at high precipitation thresholds. It is far from surprising that calibration performance is closely linked to model performance, but even the monthly rank of the raw ensemble tends to parallel the rank of the subsequent calibration closely. For example, there is only one grid cell in Table 1 where the RSM and NET ranks differ by more than one (1200 UTC March, lower Colorado), and half of eight sequence pairs are within one permutation of being the same rank. Calibrations for other precipitation thresholds exhibit behavior that

is similar to those just noted for 10 mm (results not shown).

Precipitation forecasts over the smaller watersheds that make up the larger USGS hydrologic zones provide information on performance at a spatial scale, which is arguably more germane for many hydrometeorological applications, such as driving a runoff model and providing guidance for local hydrological forecasts. For that reason, we show results for three watersheds where a relatively high number of the heaviest (50 mm) precipitation events exist in our sample: the Sacramento, San Joaquin, and Tulare Basins of the Central Valley of California (see Fig. 1). Figure 6 shows the BrSS values based on the training samples collected either from the entire USGS California region (the default local training) or from just one watershed in the Central Valley. The two calibration strategies show no significant differences except at the highest threshold of 50 mm over the Tulare Basin, where a significant improvement in BrSS occurred with the expanded training sample from the entire California region. Multiplicity considerations when conducting independent hypothesis tests (Wilks 2006, section 5.4) imply that it is highly likely that at least 1 of the 18 (nonindependent) differences (three regions, six thresholds, all correlated) should be expected to show a difference at the 10% level. In this case we believe that there is a sound reason to anticipate that higher BrSS values represent more than a sampling fluctuation. The Tulare basin is the driest of the three basins and contains far fewer heavy events than do either the Sacramento or San Joaquin basins. We postulate that the inclusion of only heavy events in the training sample from regions with similar synoptic climatologies (mountain precipitation from upslope maritime flow) may have supplied the neural network with only properly conditioned "hit" events to allow convergence toward a stable, more accurate calibration. It is clear that restricting the training sample to regions that indeed have a similar synoptic climatology, as opposed to defining the training based on the watersheds, could lead to much better performance. Further sensitivity tests are warranted.

d. Finescale, spatial distribution of skill

Figure 7 presents the spatial distribution of the ranked probability skill score (RPSS; Wilks 2006) at each stage IV pixel for the 5 months. The RPSS is an extension of the BrSS to mutually exclusive, collectively exhaustive (MECE) multiple categories. Here, we use the four lower boundaries of 1, 10, 25, and 50 mm to define five MECE categories for the RPSS estimate. Regions of positive skill are confined to coastal

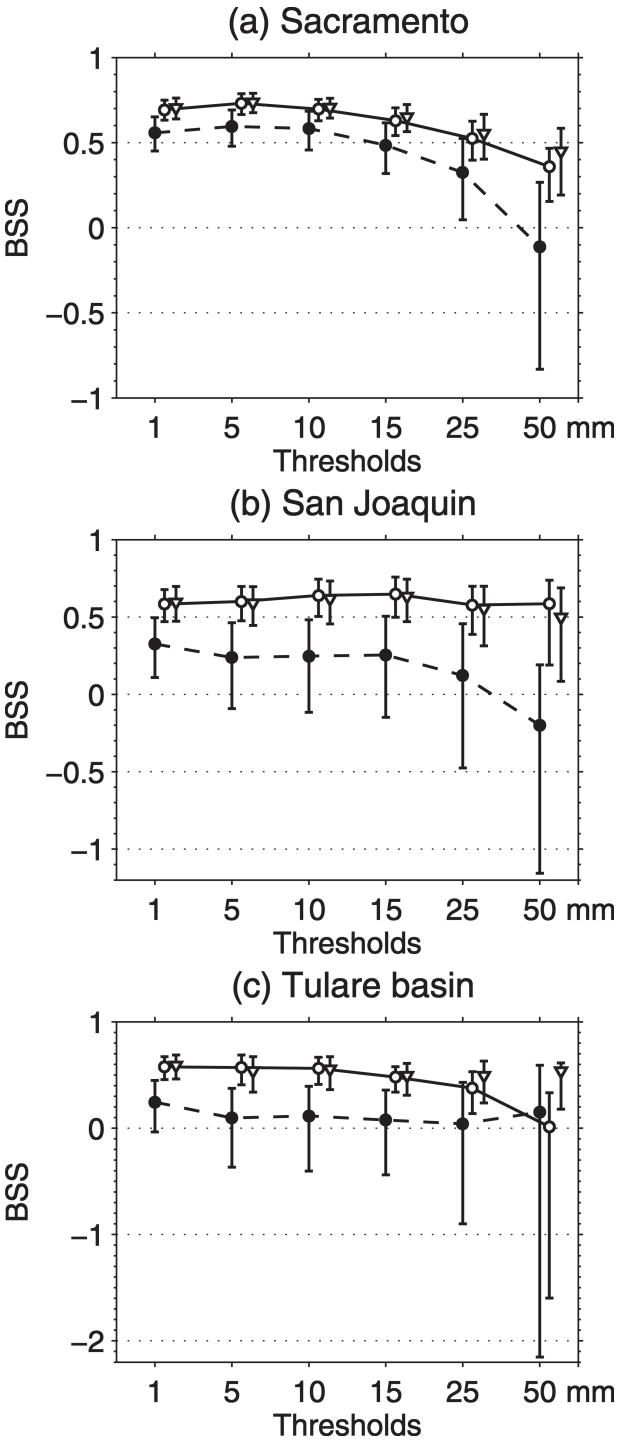

FIG. 6. Brier skill scores for the RSM forecasts (dashed line with black circles) and the NET calibrated forecasts (solid line with open circles) over the three watersheds of the Central Valley: (a) Sacramento, (b) San Joaquin, and (c) the Tulare basin. Error bars indicate 90% confidence bounds. Skill for calibration based on a larger training dataset that covers the entire CA region is indicated by open triangles. The NET curves are slightly offset to the right of the RSM curves for clarity.

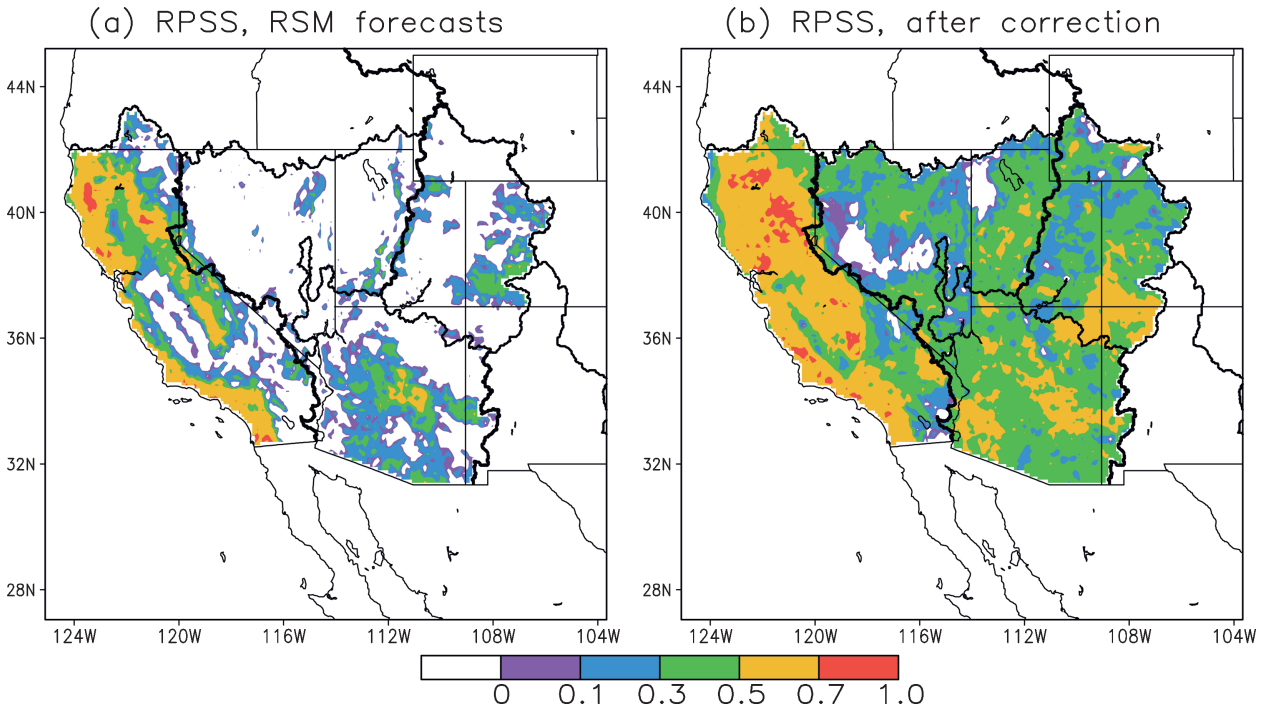

FIG. 7. Distribution of the ranked probability skill scores for the 0000 UTC RSM forecast cycle (left) before calibration and (right) after calibration. Skill is computed from five mutually exclusive, collectively exhaustive categories defined from boundaries at 1, 10, 25, and 50 mm day^{-1}. Boundaries are shown for the four USGS hydrologic unit regions.

California, the western slopes of the Sierra Nevada, Arizona, and some mountain crests over the two Colorado basins prior to calibration (Fig. 7a); in an aggregate sense, less than a half of the pixels over the four regions show skill. The total number of skillful pixels greatly increases after calibration (Fig. 7b), so much so that fewer than 5% of the pixels (most of which are situated over the Great Basin) lack skill. The neural network also improves the level of skill in most (but not all) regions, and in some locations the improvement is spectacular, such as in southwestern Arizona where the RPSS increases to 0.5 or higher. Perhaps the most serious degradation occurs along the immediate coastline of northern California, where the RPSS drops to less than 0.5. Overall, the calibration increases skill in nearly all regions, and it transforms vast areas of the interior without skill to zones with at least marginal or even moderate skill.

It is of interest to examine a representative day to illustrate some of the typical changes that the NET calibration makes to the unaltered RSM probabilities that, over the 5-month period, improve forecast skill. Figure 8 shows the distribution of PQPFs for six thresholds (1, 5, 10, 15, 25, and 50 mm day^{-1}) before (middle column) and after (right column) calibration for 24-h precipitation ending at 0000 UTC 9 November 2002. This period contains the first winter storm event in our sample that produced heavy precipitation ($\geq$50 mm day^{-1}) over a large region of California. Inspection of the RSM (middle) and NET (right) columns in Fig. 8 indicates that the calibration systematically reduces the probabilities over all regions and across all thresholds.

A comparison of the stage IV coverage for each threshold (Fig. 8, left column) and the forecast fields reveals that the effectiveness of the calibration varies by region and threshold. Consider the two lowest thresholds (1 and 5 mm). The unaltered RSM probabilities, when area accumulated over watersheds, exceed the stage IV counterparts, or $\Sigma_A \hat{p}_i(\text{RSM}) > \Sigma_A o_i$, where A is the verification domain. The wet bias is particularly acute for catchments over the Great Basin and the lower and upper Colorado basins. After calibration, the accumulations are much closer to equal ($\Sigma_A \hat{p}_i(\text{NET}) \approx \Sigma_A o_i$); thus, the NET mitigates the conditional bias, which leads to a decrease in the reliability term and an improvement in Brier skill. If we consider the two highest amounts (25 and 50 mm), however, we find that NET degrades the RSM forecast over California. The stage IV analysis shows that 50-mm rainfalls, arguably an amount of greatest hydrological concern, occur in a ring about the Central Valley. The overall

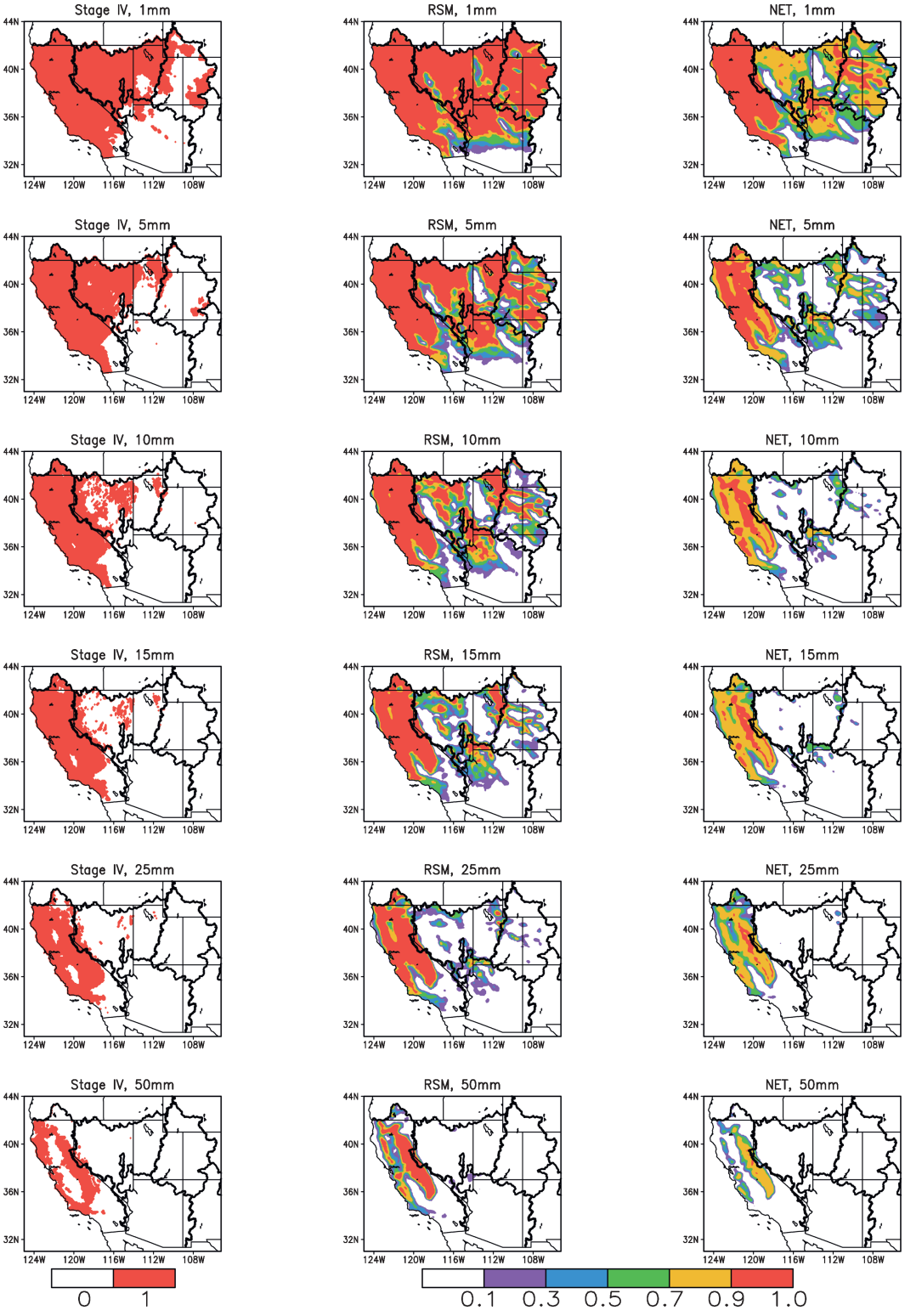

FIG. 8. Distribution of (left) stage IV precipitation estimates, (middle) RSM probabilities, and (right) the calibrated probabilities for 24-h precipitation ending at 0000 UTC 9 Nov 2002. Shading in the left column denotes pixels where the stage IV estimate exceeds the indicated threshold. Shading in the middle and right columns indicates forecast probabilities for exceeding the indicated threshold. Boundaries are shown for the four USGS hydrologic unit regions.

pattern and size of nonzero probabilities from the RSM coincide better with the stage IV estimate than the NET forecast, which for this case produces too much drying ($\Sigma_A \hat{p}_i(\text{NET}) < \Sigma_A \hat{p}_i(\text{RSM}) \sim \Sigma_A o_i$). Excessive drying also characterizes the 25-mm threshold over California, but on the positive side, the NET reduces the glut of spurious nonzero probabilities that exists in the RSM ensemble over the high terrain of the three interior basins.

The 24-h period ending at 0000 UTC 9 November 2002 yields results that are consistent with the 5-month-average results for low and moderate rainfalls; this is a substantial improvement in terms of reliability with a small or insignificant loss in discrimination ability (i.e., resolution term). Outside of the systemic reduction in probabilities that marks *every* NET forecast, the 9 November 2002 results for the upper thresholds are more equivocal and consistent with the notion of greater day-to-day volatility in the improvement provided by the NET.

5. Summary

Mitigation of state-dependent and parameter-dependent biases through statistical postprocessing will remain a problem of critical importance as long as prediction systems produce forecasts with intolerable errors. This is clearly the case for precipitation, arguably the most difficult weather element to forecast accurately (e.g., Olson et al. 1995), for the foreseeable future. In this paper, we assessed the ability of a feed-forward neural network to calibrate 24-h probabilistic quantitative precipitation forecasts from a 12-km version of the NCEP RSM ensemble forecast system over the southwest United States during the 2002–2003 cool season. Verification was performed on the stage IV mesh ($\sim$4 km $\times$ $\sim$4 km), a sufficiently fine resolution to be of hydrologic relevance in the mountainous terrain of the Southwest. Cross validation was used for training the neural network, and nonparametric bootstrapping was used to estimate the confidence bounds of the results.

The calibration procedure results in a significant increase in model skill measured relative to the sample frequency and a reduction in the day-to-day variability of the forecast skill. The improvement in the BrSS and RPSS comes from a systematic shift in the forecast distribution from the high-threshold and high-probability categories to the lower ones that reduce a conditional wet bias and the associated reliability term. The reduction comes at the expense of the resolution term as the forecast distribution is pushed closer toward the climatologic frequency. The trade-off is particularly large over regions where event occurrence in the verifying/training sample is rare.

A lingering question is how to postprocess ensemble forecasts without producing an intolerable loss in the ability to discriminate the event. Recent results by Hamill et al. (2006) show that it is possible to improve both the reliability term and the resolution term significantly. The NET was able to improve skill without a significant decrease in resolution, but only for low thresholds over California, a region where the input RSM forecasts are more skillful and rainfall events occur frequently. There are several fundamental differences between the two studies, besides different calibration methodologies, that may account for the different outcomes. Two important ones are the sample size used for the training and resolution of the verification data. Hamill et al. (2006) used a $\sim$25 yr record of global ensemble reforecasts for their training, which provides the opportunity to sample many more "rare" events during the training than was possible for our 4-month training sample. They also used precipitation analyses from the 32-km North American Regional Reanalysis (NARR) as truth for their verification. Use of the coarser NARR grid would exclude the much finer scales of the 4-km stage IV mesh, or those scales that are the most unpredictable.

We believe that the neural network approach of this study, if trained by a longer historical record that is confined to a region with a very similar climatology, is capable of producing PQPF calibrations that could improve, or at worse would not degrade, the resolution term at the 4-km scale. Further enhancement is likely, at least early in the forecast when precipitation is most predictable, if the training is multivariate. The performance of model output statistics (MOS; Glahn and Lowry 1972) clearly shows that additional predictors besides model precipitation, each containing some independent information related to the predictand, can sharpen the discrimination ability. Examples of those predictors are precipitable water, relative humidity, vertical velocity, and convective indices. Multivariate techniques would likely require longer training periods than what we now have available, and hence they were not attempted in this study. Of course, any calibration will be always subject to the limitations imposed by the time it takes for the model to reach nonlinear saturation and the inherent predictability of the atmospheric phenomenon itself.

Acknowledgments. The authors graciously acknowledge the support of NASA EOS-IDS Grant NAG5-3460 and the NSF STC program (Agreement EAR-9876800). The third author (SLM) received partial sup-

port from ONR N00014-99-1-0181, NSF ATM-0135801, and NSF ATM-0432232. Computer resources were obtained under the support of ONR N00014-00-1-0613. Special thanks are extended to Dr. Kuo-lin Hsu for providing the neural network software and Ms. Annie Reiser for editing the manuscript. We also thank the two anonymous reviewers for their insightful reviews.

REFERENCES

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.,* **131,** 1509–1523.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.,* **133,** 1076–1097.

Du, J., and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: An update. Preprints, *Ninth Conf. on Mesoscale Processes,* Fort Lauderdale, FL, Amer. Meteor. Soc., 355–356.

——, J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New Dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members. Preprint, *WMO Expert Team Meeting on Ensemble Prediction System,* Exeter, United Kingdom, WMO. [Available online at http://www.emc.ncep.noaa.gov/mmb/SREF/reference.html.]

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting,* **13,** 1132–1147.

——, and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting,* **20,** 328–350.

Eckert, P., D. Cattani, and J. Ambhul, 1996: Classification of ensemble forecasts by means of an artificial neural network. *Meteor. Appl.,* **3,** 169–178.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman and Hall, 436 pp.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Gneiting, T., and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science,* **310,** 248–249.

Hall, T., H. E. Brooks, and C. A. Doswell III, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting,* **14,** 338–345.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.,* **81,** 2653–2664.

——, J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

Hsieh, W. W., and B. Tang, 1998: Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Amer. Meteor. Soc.,* **79,** 1855–1870.

Hsu, K., H. V. Gupta, and S. Sorooshian, 1995: Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.,* **31,** 2517–2530.

——, X. Gao, S. Sorooshian, and H. V. Gupta, 1997: Precipitation estimation from remotely sensed information using artificial neural networks. *J. Appl. Meteor.,* **36,** 1176–1190.

——, H. V. Gupta, X. Gao, and S. Sorooshian, 1999: Estimation of physical variables from multichannel remotely sensed imagery using a neural network: Application to rainfall estimation. *Water Resour. Res.,* **35,** 1605–1618.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Wiley and Sons, 240 pp.

Juang, H.-M. H., and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.,* **122,** 3–26.

Koizumi, K., 1999: An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network. *Wea. Forecasting,* **14,** 109–118.

Kretzschmar, R., P. Eckert, D. Cattani, and F. Eggimann, 2004: Neural network classifiers for local wind prediction. *J. Appl. Meteor.,* **43,** 727–738.

Kuligowski, R. J., and A. P. Barros, 1998: Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Wea. Forecasting,* **13,** 1194–1204.

Manzato, A., 2005: The use of sounding-derived indices for a neural network short-term thunderstorm forecast. *Wea. Forecasting,* **20,** 896–917.

Marzban, C., 2003: Neural networks for postprocessing model output: ARPS. *Mon. Wea. Rev.,* **131,** 1103–1111.

Mass, C. F., 2003: IFPS and the future of the National Weather Service. *Wea. Forecasting,* **18,** 75–79.

Mullen, S. L., and R. Buizza, 2004: Calibration of probabilistic precipitation forecasts from the ECMWF EPS by an artificial neural network. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences,* Seattle, WA, Amer. Meteor. Soc., J5.6.

——, M. Poulton, H. E. Brooks, and T. M. Hamill, 1998: Postprocessing of ETA/RSM ensemble precipitation forecasts by a neural network. Preprints, *First Conf. on Artificial Intelligence,* Phoenix, AZ, Amer. Meteor. Soc., J31–J33.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

Murphy, S. L., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting,* **10,** 498–511.

Roebber, P. J., S. L. Bruening, D. M. Schultz, and J. V. Cortinas Jr., 2003: Improving snowfall forecasting by diagnosing snow density. *Wea. Forecasting,* **18,** 264–287.

Scherrer, S. C., C. Appenzeller, P. Eckert, and D. Cattani, 2004: Analysis of the spread–skill relations using the ECMWF Ensemble Prediction System over Europe. *Wea. Forecasting,* **19,** 552–565.

Stensrud, D. J., J. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.,* **128,** 2077–2107.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Tracton, M. S., and J. Du, 2001: Short-range ensemble forecasting (SREF) at the National Centers for Environmental Prediction. WMO Ensemble Expert Meeting Lecture 11, 11 pp. [Available online at http://www.emc.ncep.noaa.gov/mmb/SREF/Tracton_Du.forWMO2001.doc.]

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2d ed. Elsevier, 627 pp.

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Wea. Rev.,* **133,** 279–294.